Background and Goals for Fellowship Training

A Doctoral Dissertation and Research Experience

A.1 Undergraduate Research

A.1.1 Research Intern (2010-2012)

Clinical Outcomes Research Office, Washington University at St. Louis School of Medicine

Advisors: Dr. Jay Piccirillo, clinician and professor & Dr. Dorina Kallogjeri, clinician and biostatistician

Project 1: I developed evidence-based prognostic models for multiple cancer types based on patient-specific data from the Barnes Jewish Hospital oncology registry. I used methods including (1) recursive partitioning to classify observations into groups and predict within-group survival, (2) Nomograms to graphically represent the contributions to a predictive model for prognosis and (3) Cox proportional hazards models to estimate prognosis using regression. The prognostic models were validated and updated to include conditional survival based on cancer treatment options. This research is being patented and was implemented in the form of an online tool.

Project 2: I analyzed comorbid condition scoring methods for cancer patients. This research evaluated the primary methods for scoring comorbidity that impact patient's survival. In order to analyze scoring methods, I prepared cancer datasets obtained from nine hospitals. I classified each cancer case with ICD-9 coding and calculated the ACE-27 and Charlson comorbidity score for each observation. I tested for differences in survival by score to see how effectively the comorbidity score distinguished survival probabilities. This analysis showed that the scoring methods did not distinguish between survival rates equally, and the results were reported in a published manuscript.

A.1.2 Honors Undergraduate Researcher (2011-2013)

Department of Biostatistics, Gillings School of Global Public Health at the University of North Carolina

Advisor: Dr. Eric Bair, research assistant professor

Project: I proposed a novel clustering algorithm. Clustering methods allow data to be partitioned into homogeneous groups in order to identify subsets within the data, such as disease subtypes, which may be secondary or outcome-related. This is especially useful for highly heterogeneous conditions. Conventional clustering methods generally do not produce clusters that are related to an outcome of interest, particular when there are many high-variance features in the data. I developed a clustering method based on sparse clustering to improve upon existing methods by identifying clusters that are secondary or associated with a biological outcome variable. The method was validated with extensive simulation studies and implemented the procedure as a package in R. These sparse clustering methods are pertinent to many diseases because identifying clinically relevant subtypes could allow for more accurate diagnosis of disease and more personalized medicine. I further collaborated with the principle investigators of the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study, a large prospective cohort study of temporomandibular disorder (TMD), which is a form of chronic orofacial pain. We applied clustering methods to identify subtypes of TMD. This work resulted in two manuscripts; an article describing the methodology and a second article applying the methodology to OPPERA data.

A.1.3 Summer Undergraduate Research Fellow (2012)

Department of Biostatistics, Gillings School of Global Public Health at the University of North Carolina

Advisor: Dr. Eric Bair, research assistant professor

Project: I was awarded a Summer Undergraduate Research Fellowship to study the association between orofacial pain and the use of oral contraceptives (OCPs). Previous studies have shown an association between OCP use and orofacial pain, but the results have been inconsistent as to the direction of the effect. I analyzed this association by first evaluating the relationship between OCP use and pain in various body regions using mixed effects models. This showed that OCP use was associated with pain isolated to the head and face regions. I performed further analyses, assessing the longitudinal relationship between stopping and starting contraceptive use and reporting pain. I prepared a poster describing these results that won the Best Poster award at the 2012 meeting for the International Pelvic Pain Society, and a manuscript describing these results is in preparation.

A.2 Graduate Research

A.2.1 Rotation Student (2014)

Department of Biomedical Informatics, Harvard Medical School

Advisor: Dr. Peter Park, associate professor

Project: The primary focus of my rotation was to compare copy number calling methods across various geno-

typing platforms. This type of genetic data has been shown to be associated with disease phenotypes and is obtained via array or sequencing. These two different platforms, each of which have a variety of methods for determining variants from raw data, can report differences in variant calls. I sought to characterize overall trends and differences by using data on kidney renal clear cell carcinoma and colorectal cancers from The Cancer Genome Atlas. I studied differences in the sizes of called copy number variants and locations on the genome, and summarized the observations in a manuscript that is in preparation.

A.2.2 Visiting Researcher (2015-2016)

Behavioral Science Research, Boston University Goldman School of Dental Medicine

Advisor: Dr. Belinda Borelli, professor

Project: I was senior statistician on a study investigating the subtypes of smokers who are unmotivated to quit. Subtype identification could provide insights on motivating treatment entry and developing treatment content for unmotivated smokers. I performed latent class analysis on a study of 500 smokers unmotivated to quit within thirty days. I identified three distinct subtypes, which we labeled as "smokers with psychosocial barriers," "health-concerned smokers" and "unconvinced smokers." I confirmed the findings through a statistical validation study. I performed all statistical analyses and contributed to a manuscript that is currently under review.

A.2.3 Visiting Researcher (2016-2017)

Neurobiology of Fear Lab, McLean Hospital and Harvard Medical School

Advisors: Dr. Guia Guffanti Masetti, associate neuroscientist & Dr. Kerry Ressler, professor and chief scientific officer

Project: I am working on a project on the genetics of psychological disorders, which have been particularly challenged by the weakness of sparse effects across the genome. I applied my mediation methodology from my dissertation research as detailed in my research strategy to the Grady Trauma Project (GTP), a cross-sectional study of post-traumatic stress disorder (PTSD). Individuals respond differently to traumatic stress and it has thus been suggested to have a genetic or genomic component. I have been leveraging the GTP data to identify biological mechanisms and mutations via mediation analysis in order to better biologically characterize the contributions of genetics and genomics to PTSD phenotypes.

A.2.4 Doctoral Dissertation Researcher (2015-2018)

Department of Biostatistics, Harvard T.H. Chan School of Public Health

Advisors: Dr. Xihong Lin, professor and chair, Dr. John Quackenbush, professor

Project: My exposure to high dimensional data analysis during research for my undergraduate honors thesis motivated me to continue pursuing statistics for "big data" problems. I am currently focusing on the development of methods for integrating data from multiple genetics and genomics sources to learn about the biomolecular features of disease.

For instance, diseases of the lung can be complex and involve numerous genetic, epigenetic and other genomic factors. Many existing analytical methods have used simple associations between these factors which fails to capture the complexity of the disease. I am working on statistical methods to analyze this complex set of relationships while utilizing an understanding of biological processes. Thus my methods should be grounded in a correct representation of the biological mechanisms and integrating all features of the data. This work should integrate the data to provide insight into the underlying biology of lung diseases such as chronic obstructive pulmonary disease (COPD) and asthma, and features that can be intervened upon to treat and manage disease.

For the first component of my thesis work, I have proposed a mediation method to allow for the analysis between an exposure, mediator and outcome when the outcome is common and binary. Existing methodology for estimating mediation relationships for binary outcomes is limited to rare outcomes without imposing additional assumptions. Preliminary results suggest that my proposed method performs well for outcomes as prevalent as 20-70%. To examine the utility of this method in practice, I will apply this approach to data from the Normative Aging Study. I will assess whether methylation mediates the relationship of genotype on diagnosis of asthma.

I am also interested in representing the relationships of genetics and genomics using network models. Networks are a natural way to represent complex data by formalizing dependencies between different data points. We expect that the structure and properties of biological networks can provide insight on the factors that drive regulatory networks and disease development. Existing methods for networks measure a variety of metrics but do not include techniques for estimating their error. Error estimation is critical for assessing the reproducibility of constructed networks and expressing confidence in metrics. To address this gap in the literature, I have proposed

a nonparametric method for estimating the error of network metrics of interest. This method will be applied to an eQTL network of COPD. Coupled with the mediation methodology, I will be able to estimate scientifically hypothesized mediation and network relationships between multiple genomic features and assess their impact on disease.

B Training Goal and Objectives

My research plan allows me to obtain the essential skills for achieving my ultimate career goal of a professorship in biostatistics. My general research interests lie in elucidating biomedical pathways that can be intervened on by performing integrative genomic analysis. I believe that in order to obtain genomic analysis results that are informative, we must develop methods that directly represent a hypothesized relationship between features and use all available information. I have a solid foundation of statistical knowledge and experience in research which will aid in my transition to independent research. My goal for this fellowship training is to obtain the ability to develop novel statistical methods for important biomedical questions and the capacity to build meaningful collaborations with statisticians, scientists and clinicians. I am particularly interested in academia as it allows one to work highly collaboratively and as I woman in science I have the opportunity to promote diversity in the field in future students. The training plan developed, with the support of Dr. Lin and Dr. Quackenbush, works towards the following goals:

B.1 Develop a toolbox of statistical methodology.

I have completed the coursework for a doctorate in biostatistics, including general statistical theory and applied courses on statistical genetics. I will continue building my statistical knowledge by reading, implementing research and dealing with various data types. I will keep broadening my knowledge of statistical approaches by attending seminars and workshops hosted by our department. Additionally, I will continue to have informal discussions with my fellow students to think through statistical problems.

B.2 Learn to fully utilize high performance computing techniques.

Biostatistics relies on statistical programming for its implementation, and I have been using R since my undergraduate research. As I work on developing my own approaches, I will continue to formulate each distinct method as R packages for open sharing. I will also attend computing short courses at Harvard Medical School and other institutions as available. This will be important to ensuring that my methods are able to be broadly used, given that genomics is an inherent big data problem.

B.3 Gain experience in collaborating with clinicians and scientists.

Collaborations are an essential part of the biostatistical experience. Working with others provides scientific and clinical knowledge of diseases, ensuring that methods have a true impact. I will continue to attend the COPD working group and meet with Dr. Silverman to learn about COPD and improve my biological understanding of lung diseases and their genetic underpinnings. I will connect with individuals who have successfully built research careers to learn the skills necessary for making novel scientific discoveries with adept scientists.

B.4 Develop communication and presentation skills.

I will continue to develop my skills in explaining high level statistical topics by presenting a formal research talk twice a semester at Dr. Lin's group research meeting. At Dr. Quackenbush's more informal weekly group meeting, I will aim to present recent results and discuss my work at least every other week. I will present my research at the Department of Biostatistics' Student Seminar, the COPD working group and larger local meetings. I will further pursue presenting at statistics and biomedical conferences locally and nationally to improve my speaking skills. If awarded, this grant will provide the support necessary to attend such conferences.

B.5 Publish in peer-reviewed scientific journals.

I plan to publish each of the aims of this proposal, in addition to my other research in progress, in peer-reviewed journals. I will participate in regular journal clubs and manuscript preparation to supplement my knowledge of the publishing process and best writing practices. I will make myself available to assist my advisors in their roles as reviewers of articles by reading and critiquing journal submissions.

B.6 Gain teaching and mentorship experience.

I have experience in explaining statistical concepts to those with minimal exposure to statistics through classroom teaching, tutoring and working with clinicians without quantitative training. In order to succeed as an independent investigator, one must be capable of teaching and mentoring students. I will continue to tutor small groups and individuals throughout my degree as I enjoy sharing my statistical knowledge and particularly helping others learn basic statistics, to code, and to fully interpret research findings.

B.7 Promote women and minorities in science and statistics.

As a woman coming from a mathematics background, I have experienced the value in having strong female mentorship as the minority in the field. I will actively encourage my fellow female scientists and continue to serve as the biostatistics and epidemiology tutor for the Commonwealth Fellowship, an esteemed diversity program training clinicians in public health.

C Activities Planned Under this Award

Year	Activity	Percent Time
1	Research	80
	Professional Development	5
	Workshops & Conferences	5
	Clinical Collaborations	10

This allocation of my year will ensure that I obtain the statistical and scientific skills to undertake novel research and develop my professional skill set, such as effective scientific communication.

C.1 Research

The primary focus in my research is to develop statistical theory to model genomic data and bolster my computing skills in order to implement methods on this big data. I will also need to stay up to date on the literature in the area of integrative genomics. I will continue to meet biweekly with Dr. Lin and with Dr. Quackenbush. Additionally, I will continue having joint meetings with Drs. Lin and Quackenbush, and hold them every quarter. I will also continue to attend Dr. Lin's biweekly group research meeting to present research, contribute to discussion on statistical methods and learn about recent statistical developments through journal presentations. Dr. Lin is a recognized leader in statistical genetics and provides support on statistical theory. I will maintain attendance at Dr. Quackenbush's weekly lab meeting, where we have open discussion of research in progress to receive and contribute feedback in an interdisciplinary group. Dr. Quackenbush complements the statistically theoretical support with demonstrated knowledge in computational biology, specifically algorithms and applied methods for answering scientific questions of interest. I will keep attending Dr. Silverman's weekly group meeting on COPD projects. This will provide information on the research methods particular to this biomedical area and allow for greater understanding of lung diseases. These group meetings will allow for practice presenting research and improve my ability to assess and critique others' research. I will gain exposure to issues in statistics, computational biology and methods applied specifically to COPD.

C.2 Professional Development

This activity includes coordinating and attending local seminar series as well as staying involved in activities to improve "soft skills". I am the student coordinator of the Biostatistics & Biomedical Informatics Big Data (B3D) seminar hosted by the Biostatistics and Biomedical Informatics departments at Harvard University. The B3D seminar features research talks from leaders in statistical, computational and machine learning methods for analyzing complex biomedical data. I also attend my departmental Student Seminar series to connect with other students and discuss current research and computing approaches. There are many other aspects of development that are critical to academic success, such as grant writing, manuscript writing, and public speaking. I will attend my department's Career Development Series which instructs on these topics. Our school as well as other institutions offer trainings across these skills and, if awarded, this grant could fund additional training. For instance, I could attend the "NIH Perspectives on Peer Review" tutorial or participate in the "Publishing Without Perishing: Strategies for Success in Publishing in Biostatistical Journals" roundtable discussion with biostatistics leader Dr. Marie Davidian at the ENAR Spring Meeting.

C.3 Workshops & Conferences

I have found attending conferences to be important to gaining exposure to the most recent biostatistical methods and areas of application. If awarded, this grant would help to fund attendance at conferences relevant to my research. I would like to be able to present my statistical methods at the Joint Statistical Meetings, where there is the opportunity to discuss rigorous methodology. I would also like to present the applications of my methods at the American Society for Human Genetics conference, which provides insight to the status and direction of the field and allows for many interdisciplinary conversations.

Computing is an increasingly essential component of biostatistics research, thus it is important to maintain up-to-date knowledge of high performance computing. In June 2016, I attended the International High Performance Computing Summer School to gain a more comprehensive understanding of computational methods to perform intensive, high-level analyses. This experience has already proved important for leveraging the high-dimensional data that I study; therefore, this grant would allow for funding to pursue further computing instruction through workshops and meeting such as the International Conference for High Performance Computing, Networking, Storage and Analysis and Open Source and the Interactive Scientific Computing with R/Jupyter tutorial at the ENAR Spring Meeting.

C.4 Teaching

This activity has included teaching as a teaching assistant in a classroom setting as well as a private and fellowship program tutor. This has required the development of course materials, holding office hours, and instruction. These teaching responsibilities prepare me for future teaching roles and the communication of statistical concepts and methodology to those who do not have a quantitative background. I plan to continue tutoring for the Commonwealth Fellowship Program at the Harvard Chan School, which includes tutoring on introductory statistics and epidemiology to medical doctors.

C.5 Clinical Collaborations

This activity includes meetings with Dr. Ed Silverman for guidance on clinical and biological interpretations of COPD disorder. We will discuss the biomolecular basis of disease and aspects of the data that has been collected throughout the study. It further includes meetings with other collaborators, such as Dr. Kerry Ressler and Dr. Guia Guffanti at McLean Hospital. This is in continuation of collaborations that allow for access to interesting data and scientific questions for applications of developed statistical methods.